

## Phylogenetic Profilers

Genes that may require further analysis and functional annotation can be identified using comparative analysis tools within a specific genomic context. Phylogenetic profiles are designed for such purpose.

### Single Genes

In many cases the differences in physiology, phenotypic properties and ecology of different organisms can be attributed to the differences in their gene content, i.e., the differences in abundance of various gene families, including the ultimate case of certain genes being present in one genome but not in another genome(s) and vice versa. Therefore the genes identified as more or less abundant (or present or absent) when comparing the genome of interest to its genome context, often become the focus of microbial genome analysis and may require special attention from the annotator.

The Phylogenetic Profiler for Single Genes tool allows finding genes in a specific genome that have / do not have homologs in other related genomes. There are two steps involved in such a selection:

(a) Start with Find Genes in the Main Menu and select Phylogenetic Profiler for Single Genes under the Phylogenetic Profilers second level menu bar.

(b) Select a target genome and set the condition for selecting its genes with respect to presence or absence of homologs in other related genomes.

(c) (Optional) Additional features are available for setting cutoffs and adding extra fields in the output display.

**Phylogenetic Profiler for Single Genes**

Sequencing Status: Finished Domain: Archaea

Selected Genomes: Thermoplasma volcanium GSS1 (A) [F]

Search for: <enter a genome name to search>

Thermococcus gammatolerans E33 (A) [F]  
Thermococcus guaymasensis DSM 11113 (A) [F]  
Thermococcus kodakarensis KOD1 (A) [F]  
Thermococcus litoralis DSM 5473 (A) [F]  
Thermococcus nautoli 30-1 (unscreened) (A) [F]  
Thermococcus onnurineus NA1 (A) [F]  
Thermococcus peptonophilus DG-1 (A) [F]  
Thermococcus piezophilus CDGS (A) [F]  
Thermococcus sibiricus MM 739 (A) [F]  
Thermococcus sp. 2319x1 (A) [F]  
Thermococcus sp. 4557 (A) [F]  
Thermococcus sp. A501 (A) [F]  
Thermococcus sp. AM (A) [F]  
Thermococcus sp. E51 (A) [F]  
Thermofilum carboxydiphilus 1505 (A) [F]  
Thermofilum pendens Hrk 5 (A) [F]

**Summary Statistics**

Feature	Number	Gene Number	% of Total
Total number of genes	276	100.00%	

Missing Gene? Tblastn of the first selected gene in the list below against the genomes selected in Without Homologs In Genomes.

Add Selected to Gene Cart Select All Clear All

Filter column: Result Filter: text Apply

Export Page 1 of 3 << first < prev 1 2 3 next > last >> 100

Select	Result	Gene ID	Locus Tag	Gene Name	Length
<input type="checkbox"/>	1	638190495	TVG0007048	putative transposase	168
<input type="checkbox"/>	2	638190515	TVG0025913	arsenite efflux ATP-binding protein ArsA (TC 3.A.4.1.1)	258
<input type="checkbox"/>	3	638190532	TVG0045165	transposase	188
<input type="checkbox"/>	4	638190537	TVG0049983	H <sup>+</sup> -transporting ATP synthase subunit K	75
<input type="checkbox"/>	5	638190559	TVG0067031	hypothetic protein	169

**Figure 1.** Find genes with the Phylogenetic Profiler for Single Genes tool.

**Example 1.** After setting the genome context to two genomes, *T. volcanium* (*Thermoplasma volcanium* GSS1) and *T. acidophilum* (*Thermoplasma acidophilum* DSM 1728), use the Phylogenetic Profiler to find *T. volcanium* genes that have no homologs in *T. acidophilum*, as shown in Figure 1(i). Similarity cutoffs can be used to fine-tune the selection. To choose your own cutoff values, scroll down to find **Similarity Cutoffs** under **Advance Options**. The list of genes with the specified profile are then provided as a selectable list as shown in Figure 1(ii). Users can also select to add additional fields such as COG, Pfam, etc. in the result display. Such selections are available in the **Function Display Options** under the same **Advance Options** in the window shown in Figure 1(i).

The Phylogenetic Profiler for Single Genes can be used, for example for finding unique, conserved, or gained genes in the target genome with respect to other genomes of interest. In the example shown in Figure 1, 276 genes are found to be unique in *T. volcanium* with respect to *T. acidophilum*. These genes can be selected to add into the Gene Cart for further analysis.

## Gene Cassettes

The Phylogenetic Profiler for Gene Cassettes allows selecting genes that are part of a gene cassette (i.e., are collocated on the chromosome) in a query genome and are part of related (conserved part of) gene

cassettes in other genomes. (**Limitation:** Currently you can only select up to **50 Collocated In Genomes.**)

To use this tool, first, select the protein cluster you wish to use in the comparison (COG or Pfam). Then select your query genome into the "Find Genes In" column, and select (up to 50) genomes for gene cassette comparisons with the query genome in the "Collocated In" list. Click "Submit" button to start the analysis.

**Phylogenetic Profiler for Gene Cassettes**

Find genes in a query genome, that are collocated in the query genome as well as across other genomes of interest, based on their inclusion in cassettes.  
**Limitation:** Currently you can only select up to 50 Collocated In Genomes.  
 Powered by [Fastbit](#)

Genome Completion: [F]inished, [P]ermanent Draft, [D]raft.

Select Protein Cluster  
☒ COG  
☐ Pfam

**Sequencing Status** **Domain**

**Selected Genomes**

Find Genes In  
 Thermoplasma volcanium GSS1 (A) [F]  
 Add Remove

Collocated In (50 max.) 1  
 Thermoplasma acidophilum DSM 1728 (A) [F]  
 Add Remove

**Submit**

Sequencing Status	Domain	Find Genes In
<input type="checkbox"/>	1 638190721 amino acid/polyamine/organocation transporter, APC superfamily (TC 2.A.3)	
<input type="checkbox"/>	2 638190724 signal sequence peptidase	
<input type="checkbox"/>	3 638190725 thioredoxin peroxidase	
<input type="checkbox"/>	4 638190726 glycosyl transferase	
<input type="checkbox"/>	5 638190728 hypothetical protein	
<input type="checkbox"/>	6 638190730 purine nucleoside phosphorylase	
<input type="checkbox"/>	7 638190731 phosphoesterase	
<input type="checkbox"/>	8 638190732 hypothetical protein	
<input type="checkbox"/>	9 638190733 hypothetical protein	
<input type="checkbox"/>	1 638190667 hypothetical protein	
<input type="checkbox"/>	2 638190678 zinc finger protein	
<input type="checkbox"/>	3 638190679 isopentenyl-diphosphate delta-isomerase (EC 5.3.3.2)	
<input type="checkbox"/>	4 638190680 isopentenyl-diphosphate delta-isomerase (EC 5.3.3.2)	
<input type="checkbox"/>	5 638190681 acetyltransferase	
<input type="checkbox"/>	6 638190682 hypothetical protein	
<input type="checkbox"/>	7 638190685 transcriptional regulator, XRE family	
<input type="checkbox"/>	8 638190686 CDP-2,3-bis-(O-geranylgeranyl)-sn-glycerol synthase	
<input type="checkbox"/>	9 638190687 peptidyl-tRNA hydrolase (EC 3.1.1.29)	

(ii) (iii) (iv)

**Figure 2.** Find genes with the Phylogenetic Profiler for Gene Cassettes tool.

**Example 2.** Suppose a user wishes to find gene cassettes in *T. volcanium* (*Thermoplasma volcanium* GSS1) collocated in *T. acidophilum* (*Thermoplasma acidophilum* DSM 1728) using COG Protein Cluster. *T. volcanium* will be added to the "Find Genes In" field, and *T. acidophilum* will be added to the "Collocated In" list as shown in Figure 2(i).

After the user clicks the Submit button, the result will show all collocated gene cassettes in both genomes as in Figure 2(ii). There are two corresponding 9-gene cassettes in *T. volcanium* (starting from gene 638190721) and in *T. acidophilum* (starting from gene 638190667), respectively. Chromosome viewer -- colored by COG for both groups are shown in Figure 2(iii) and 2(iv). (To see the chromosome viewer colored by COG, first go to the gene detail page by clicking the Gene OID, then select "COG" in

the "Chromosome Viewer colored by" dropdown list in the **Evidence For Function Prediction** section in the gene detail page.)

Note that in each specific group of collocated genes in the query genome, individual genes may correspond to parts of multiple chromosomal cassettes in the other genomes involved in the profiler condition. Users can explore the details of individual gene listed in the result by clicking on the associated Gene OIDs. Users can also select all or a subset of genes from the list to be added to the Gene Cart for further analysis.